

# Specification Tests for the Propensity Score

Pedro H. C. Sant'Anna\*

Xiaojun Song<sup>†</sup>

Vanderbilt University

Peking University

November 21, 2016

## Abstract

This paper introduces new nonparametric diagnostic tools for detecting propensity score misspecification. These tests may be applied to assess the validity of different treatment effects estimators that rely on the correct specification of the propensity score. Our tests do not suffer from the “curse of dimensionality” when the vector of covariates is of high-dimensionality, are fully data-driven, do not require tuning parameters such as bandwidths, and are able to detect a broad class of local alternatives converging to the null at the parametric rate  $n^{-1/2}$ , with  $n$  the sample size. We show that the use of an orthogonal projection on the tangent space of nuisance parameters both improves power and facilitates the simulation of critical values by means of a multiplier bootstrap procedure. The finite sample performance of the tests are examined by means of a Monte Carlo experiment and an empirical application. Open-source software is available for implementing the proposed tests.

**JEL:** C12, C31, C35, C52.

**Keywords:** Goodness-of-fit; Integrated Moments; Empirical Processes; Multiplier Bootstrap; Treatment Effects.

---

\*Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA. Email: pedro.h.santanna@vanderbilt.edu. Financial support from the Spanish Plan Nacional de I+D+I (Grant No. ECO2014-55858-P) is acknowledged.

<sup>†</sup>Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China. Email: sxj@gsm.pku.edu.cn. Financial support from the National Natural Science Foundation of China (Grant No. 71532001) is acknowledged.

# 1 Introduction

The propensity score, which is defined as the conditional probability of receiving treatment given covariates, is one of the most widely used tools in causal inference. Part of its popularity can be credited to the seminal result of [Rosenbaum and Rubin \(1983\)](#): if the treatment assignment is independent of the potential outcomes conditional on a vector of covariates, then one can obtain unbiased and consistent estimators of different treatment effect measures by adjusting for the propensity score alone, greatly reducing the dimensionality of the underlying problem. Several methods that exploits this important insight are now an essential part of the applied researcher’s toolkit. Examples includes matching, see e.g. [Rosenbaum and Rubin \(1985\)](#), [Heckman et al. \(1997\)](#), and [Abadie and Imbens \(2006, 2016\)](#); inverse probability weighting (IPW), see e.g. [Rosenbaum \(1987\)](#), [Hirano et al. \(2003\)](#) and [Donald and Hsu \(2014\)](#); regression methods, see e.g. [Hahn \(1998\)](#), [Firpo \(2007\)](#); and many others. For a review, see [Heckman and Vytlacil \(2007\)](#) and [Imbens and Wooldridge \(2009\)](#).

Despite their popularity, a main concern of these methods is that the propensity score is usually unknown, and therefore has to be estimated. Given the high dimensionality of available covariates, researchers are usually coerced to adopt a parametric model for the propensity score since nonparametric estimation methods suffer from the “curse of dimensionality”, implying that the resulting treatment effects estimator can have considerably poor properties, even for large sample sizes. Such a common practice raises the important issue of model misspecification. Indeed, as shown by [Frölich \(2004\)](#), [Kang and Schafer \(2007\)](#), [Huber et al. \(2013\)](#) and [Busso et al. \(2014\)](#), propensity score misspecifications can lead to misleading treatment effects estimates.

In this paper we propose new specification tests for parametric propensity score models. In contrast to existing proposals, our tests are fully data-driven, do not require user-chosen tuning parameters such as bandwidths, and are able to detect a broad class of local alternatives converging to the null at the parametric rate  $n^{-1/2}$ , with  $n$  the sample size. Furthermore, by exploiting the balancing property of the propensity score - i.e.

that conditioning on the true (unknown) propensity score, the treatment assignment is independent of the covariates (Rosenbaum and Rubin, 1983) our tests do not suffer from the “curse of dimensionality” when the vector of covariates is of high-dimensionality, and have dramatically greater power than competing tests for many alternatives. Of course, such power gains do not come without a cost: there exist some classes of alternative hypotheses in which our tests have trivial power. Nonetheless, we believe that such a compromise is reasonable since, as pointed out by Janssen (2000) and Escanciano (2009), achieving reasonable power over all possible directions seems hopeless.

The proposal closest to ours is Shaikh et al. (2009), who also exploits the balancing property of the propensity score. Despite using a similar characterization of  $H_0$  as Shaikh et al. (2009), our proposal greatly differ from theirs. Whereas Shaikh et al. (2009) adopts the local smoothing approach, see e.g. Hardle and Mammen (1993), Zheng (1996), Fan and Li (1996) and Li and Wang (1998), we adopt the integrated conditional moment (ICM) approach, see e.g. Bierens (1982, 1990), Bierens and Ploberger (1997), Stute (1997), and Escanciano (2006a). As a consequence, our approach inherits some advantages when compared to Shaikh et al. (2009). First, our tests do not require delicate bandwidth choices, unlike Shaikh et al. (2009) test whose performance can be very sensitive to it. Second, in contrast with Shaikh et al. (2009), our approach has power against local alternatives converging to null at the parametric rate.

Another popular procedure to assess misspecification of the propensity score model is to use “balancing score” tests. Initially proposed by Rosenbaum and Rubin (1985), these tests consist of assessing if each covariate is independent of the treatment assignment, conditional on the propensity score. This is often implemented examining whether moments (usually just the mean) of the observable characteristics between the two “matched” or “weighted” groups are the same; see e.g. Dehejia and Wahba (2002) and Smith and Todd (2005). One should bare in mind that because “balancing score” tests are usually based on a finite number of orthogonality conditions, there are uncountably many misspecification that cannot be detected with these tests. Furthermore, as shown by Lee (2013), balancing score tests tend to have severe size distortions due to the “multiple testing

problem”, and the failure to account for the estimation effect of the propensity score. Such drawbacks put at stake the reliability of many of these procedures. Our proposal, on the other hand, does not suffer from these.

Our paper also contribute to the literature on ICM tests. What appears distinctive to our approach is that (i) we exploit the dimension-reduction coming from balancing property of the propensity score, and (ii) we acknowledge our lack of knowledge of the “true” correct specification of the propensity score, by means of an orthogonal projection of certain weight-functions into the tangent space of nuisance parameters. The result of (i) and (ii) is a test with improved power properties. The power improvement due to the dimension reduction has been noticed by [Stute and Zhu \(2002\)](#), [Escanciano \(2006a\)](#) and [Shaikh et al. \(2009\)](#), whereas the power improvement due to the use of orthogonal projections has been noticed in different contexts, see e.g. [Neyman \(1959\)](#), and more recently, [Bickel et al. \(2006\)](#) and [Escanciano and Goh \(2014\)](#). To the best of our knowledge, our proposal is the first to incorporate both procedures.

The rest of the paper is organized as follows. In [Section 2](#) we present the testing framework. The asymptotic properties of our tests are established in [Section 3](#). We next examine the finite sample properties of our tests by means of a Monte Carlo study in [Section 4](#). We provide an empirical illustration of our procedures in [Section 5](#). [Section 6](#) concludes. Mathematical proofs are gathered in an appendix at the end of the article.

Finally, all proposed tests discussed in this article can be implemented via open-source R package *pstest*, which is freely available from GitHub (<https://github.com/pedrohcg/pstest>).

## 2 Testing Framework

### 2.1 Background

Let  $D$  be a binary random variable that indicates participation in the program, i.e.  $D = 1$  if the individual participates in the treatment and  $D = 0$  otherwise, and let  $X$  be an observable  $d \times 1$  vector of pre-treatment covariates. Denote the support of  $X$  by

$\mathcal{X} \subseteq \mathbb{R}^d$ . Define the propensity score  $p(x) = \mathbb{P}(D = 1 | X = x)$ . We have a random sample  $\{(D_i, X_i')'\}_{i=1}^n$  from  $(D, X)'$ . Throughout the rest of this paper, all random variables are defined on a common probability space  $(\Omega, \mathcal{A}, \mathbb{P})$ .

In this paper, we are interested in testing whether a parametric putative model for  $p(x)$  is correctly specified. Formally, we want to test

$$H_0 : \exists \theta_0 \in \Theta : E[D - q(X, \theta_0) | X] = 0 \text{ a.s.}, \quad (1)$$

where  $\Theta \subset \mathbb{R}^k$ , and  $q(X, \theta) : \mathcal{X} \times \Theta \mapsto [0, 1]$  is a family of parametric functions known up to the finite dimensional parameter  $\theta$ . Common specifications for  $q(X, \theta)$  are the Probit,  $\Phi(X'\theta)$ , and the Logit,  $\Lambda(X'\theta)$ , where  $\Phi(\cdot)$  and  $\Lambda(\cdot)$  are the normal and logistic link-function, respectively.

In order to assess the validity of (1), one can use standard nonparametric goodness-of-fit tests for regression models; see [González-Manteiga and Crujeiras \(2013\)](#) for a review. For instance, one can adopt the local smoothing approach and use [Zheng \(1996\)](#) test. Alternatively, one can adopt the ICM approach and use [Stute \(1997\)](#) test. Although both procedures would provide asymptotic valid tests under weak regularity conditions, their finite sample performance are likely to be poor when the dimensionality of  $X$  is large, as is often the case in policy evaluation applications. This problem is often referred as “the curse of dimensionality”.

To circumvent such a problem, we build on [Rosenbaum and Rubin \(1983\)](#) and [Shaikh et al. \(2009\)](#), and exploit the balancing property of the propensity score, i.e., that conditional on the propensity score  $p(x)$ , the treatment assignment  $D$  is independent of the covariates  $X$ ,

$$D \perp\!\!\!\perp X | p(X), \quad (2)$$

see [Rosenbaum and Rubin \(1983\)](#). Such important result implies that (1) can be rewritten as

$$H_0 : \exists \theta_0 \in \Theta : \mathbb{E}[D - q(X, \theta_0) | q(X, \theta_0)] = 0 \text{ a.s.} \quad (3)$$

It is important to observe that the characterization of the null hypothesis in (3) only involves a one-dimensional conditional expectation, which is in sharp contrast to (1).

As a consequence, tests based on this characterization do not suffer from the “curse of dimensionality”, and therefore enjoy better size and power properties against many alternatives. However, one must also bear in mind the “conditioning variable” in (3) is not observed, implying that tests based on this characterization must handle the “generated regressor” problem.

Shaikh et al. (2009) adapts the test proposed by Zheng (1996) to assess (3). More precisely, Shaikh et al. (2009) consider a tests statistic based on

$$\hat{V}_n(h_n) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \frac{1}{h_n} K \left( \frac{q(X_i, \hat{\theta}_n) - q(X_j, \hat{\theta}_n)}{h_n} \right) \varepsilon_i(\hat{\theta}_n) \varepsilon_j(\hat{\theta}_n), \quad (4)$$

where  $\varepsilon_i(\theta) = D_i - q(X_i, \theta)$ ,  $\hat{\theta}_n$  is a  $\sqrt{n}$ -consistent estimator of  $\theta_0$  under  $H_0$ ,  $h_n$  is a bandwidth,  $K(\cdot)$  is a kernel. Note that, in addition to the estimation of  $\theta_0$  under the  $H_0$ , Shaikh et al. (2009) procedure requires smoothing of the data, implying that its finite-sample properties relies on the adequate choice of the smoothing parameter  $h_n$ , a task that is far from trivial in testing problems.

To provide a “smoothing parameter free” testing procedure for (3), we adopt the ICM instead of the local smoothing approach. We exploit that the conditional moment restriction in (3) can be expressed as a continuum number of unconditional moment restrictions, i.e., we can re-write (3) as

$$H_0 : \exists \theta_0 \in \Theta : \mathbb{E}[\varepsilon(\theta_0) w(q(X, \theta_0), u)] = 0, \quad \forall u \in \Pi, \quad (5)$$

where  $\Pi$  is a properly chosen space and the parametric family  $\{w(\cdot, u) : u \in \Pi\}$  is such that the equivalence between (3) and (5) holds; see e.g. Bierens and Ploberger (1997) and Escanciano (2006b) for primitive conditions on the family  $w(\cdot, u)$  to satisfy this equivalence. The two most popular weighting functions are the exponential function,  $w(q, u) = \exp(iuq)$ , as in Bierens (1982, 1990), where  $i = \sqrt{-1}$  denotes the imaginary number; and the indicator function,  $w(q, u) = 1\{q \leq u\}$ , as in Stute (1997), Stute et al. (1998), Stute and Zhu (2002), Delgado and Stute (2008), among many others. Other possible weight functions include  $w(q, u) = \exp(qu)$ ,  $w(q, u) = (1 + \exp(-qu))^{-1}$ ,  $w(q, u) = \sin(qu)$ , and  $w(q, u) = \sin(qu) + \cos(qu)$ , see Stinchcombe and White (1998)

and [Escanciano \(2007\)](#).

Our tests have two main differences with respect to the standard ICM tests. First, the weight functions  $w$  depends on  $X$  only through the propensity score model under  $H_0$ , a one-dimensional (unknown) function. As a consequence, the ICM in [\(5\)](#) is insensitive to the dimension of the explanatory variables  $X$ . Second, we explicitly acknowledge that  $\theta_0$  is a nuisance parameter in testing for [\(5\)](#) by proposing to use weight functions  $w$  that leads to test statistics whose limiting distributions do not depend on the estimator used. To achieve such desirable feature, we make use of orthogonal projections on the tangent space of nuisance parameters. The result of these two features is a test with higher power properties, as is shown in [Section 4](#).

In the next subsection we describe how we construct our new projection-based tests, highlighting the role played by the orthogonal projection. It is worth stressing that our projection-based tests cover a large class of weighting functions  $w(q, u)$ .

## 2.2 The projection-based specification test

Denote  $\Pi = [0, 1]$  the unit interval. Given a random sample  $\{(D_i, X_i')'\}_{i=1}^n$ , it seems natural to construct test statistics for [\(5\)](#) based on the sample analogue

$$\hat{R}_{w,n}(u) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) w(q(X_i, \hat{\theta}_n), u), \quad (6)$$

where  $\hat{\theta}_n$  is a  $\sqrt{n}$ -consistent estimator for  $\theta_0$  under  $H_0$ . For instance, when the propensity score is assumed to be a member of the logistic (or Gaussian) family, one can estimate  $\theta_0$  by maximum likelihood (ML), non-linear least squares (NLLS), or generalized method of moments (GMM).

Tests for [\(5\)](#) can be constructed by comparing how “close” in an appropriate sense [\(6\)](#) is to zero. In the Appendix, we show that under  $H_0$  and some weak regularity conditions

provided below, the process  $\hat{R}_{w,n}(u)$  can be decomposed as

$$\hat{R}_{w,n}(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) w(q(X_i, \theta_0), u) - \sqrt{n}(\hat{\theta}_n - \theta_0)' \mathbb{E}[g(X, \theta_0) w(q(X, \theta_0), u)] + o_P(1) \quad (7)$$

uniformly in  $u \in \Pi$ , where  $g(x, \theta) = \partial q(x, \theta) / \partial \theta$ . The representation in (7) has a very important implication: whenever standard weights  $w$  such as the indicator and exponential functions are used, the asymptotic distributions of tests based on (6) are sensitive to the estimator  $\hat{\theta}_n$  used. As a consequence, for a given parametric specification  $p(x) = q(X, \theta_0)$ , the asymptotic null distributions of tests based on (7) will depend on whether one estimates  $\theta_0$  using ML, NLLS or GMM, even though the underlying specification for the propensity score is the same across these methods.

To avoid the aforementioned drawback, we consider a convenient class of  $w(q, u)$  such that

$$\mathbb{E}[g(X, \theta) w(q(X, \theta), u)] \equiv 0. \quad (8)$$

In particular, we consider a projection-based transformation of  $w(q(X, \theta), u)$ ,  $\mathcal{P}w(q(X, \theta), u)$ , where

$$\mathcal{P}w(q(X, \theta), u) \equiv w(q(X, \theta), u) - g'(X, \theta) \Delta^{-1}(\theta) G_w(u, \theta), \quad (9)$$

with

$$G_w(u, \theta) = \mathbb{E}[g(X, \theta) w(q(X, \theta), u)],$$

and

$$\Delta(\theta) = \mathbb{E}[g(X, \theta) g'(X, \theta)].$$

The intuition behind (9) is very simple. First, note that  $\Delta^{-1}(\theta) G_w(u, \theta)$  is the vector of linear projection coefficients of regressing  $w(q(X, \theta), u)$  on  $g(X, \theta)$ . Thus, it follows that  $g(X, \theta)' \Delta^{-1}(\theta) G_w(u, \theta)$  is the best linear predictor of  $w(q(X, \theta), u)$  given  $g(X, \theta)$ ,



and that (9) is nothing more than the associated projection error. It then follows that

$$\begin{aligned}\mathbb{E}[g(X, \theta) \mathcal{P}w(q(X, \theta), u)] &= \mathbb{E}[g(X, \theta) (w(q(X, \theta), u) - g'(X, \theta) \Delta^{-1}(\theta) G_w(u, \theta))] \\ &= G_w(u, \theta) - \Delta(\theta) \Delta^{-1}(\theta) G_w(u, \theta) \\ &= 0.\end{aligned}$$

Based on the aforementioned properties, our tests are based on proper continuous functionals of the (feasible) projection-based empirical process  $\hat{R}_{w,n}^p(u)$ , given by

$$\hat{R}_{w,n}^p(u) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) \mathcal{P}_n w(q(X_i, \hat{\theta}_n), u), \quad (10)$$

where  $\hat{\theta}_n$  is a  $\sqrt{n}$ -consistent estimator for  $\theta_0$  under  $H_0$ , and  $\mathcal{P}_n w(q(X, \theta), u)$  is the sample analogue of projection in (9),

$$\mathcal{P}_n w(q(X, \theta), u) = w(q(X, \theta), u) - g'(X, \theta) \Delta_n^{-1}(\theta) G_{n,w}(u, \theta), \quad (11)$$

with

$$G_{n,w}(u, \theta) = \frac{1}{n} \sum_{i=1}^n g(X_i, \theta) w(q(X_i, \theta), u)$$

and

$$\Delta_n(\theta) = \frac{1}{n} \sum_{i=1}^n g(X_i, \theta) g'(X_i, \theta).$$

Two examples that we use in the simulations and the empirical example towards the end of this paper are the following Cramér-von Mises-type and Kolmogorov-Smirnov-type functionals,

$$CvM_n = \int_{\Pi} \left| \hat{R}_{1,n}^p(u) \right|^2 F_n(du) = \frac{1}{n} \sum_{i=1}^n \left[ \hat{R}_{1,n}^p \left( q \left( X_i, \hat{\theta}_n \right) \right) \right]^2, \quad (12)$$

$$KS_n = \sup_{u \in \Pi} \left| \hat{R}_{1,n}^p(u) \right|, \quad (13)$$

respectively, where  $F_n(u) = n^{-1} \sum_{i=1}^n 1 \left( q \left( X_i, \hat{\theta}_n \right) \leq u \right)$  is the empirical distribution function (EDF) of  $q \left( X_i, \hat{\theta}_n \right)$ ,  $1 \leq i \leq n$ , and where  $\hat{R}_{1,n}^p(u)$  is defined as in (10) with  $\mathcal{P}_n w(q(X_i, \hat{\theta}_n), u) = \mathcal{P}_n 1 \left\{ q(X_i, \hat{\theta}_n) \leq u \right\}$ .

The test statistics  $CvM_n$  and  $KS_n$  should be small if the null hypothesis (3) is true, while “large” values of  $CvM_n$  or  $KS_n$  imply the rejection of  $H_0$ . Obviously, different

test statistics could be developed by applying other distances, or choosing alternative weighting functions. For ease of exposition, we concentrate on  $CvM_n$  and  $KS_n$ .

### 3 Asymptotic theory

In this section, we establish the asymptotic distribution of the projection-based empirical process  $\hat{R}_{w,n}^p(u)$  under the null hypothesis  $H_0$ , under the fixed alternative hypothesis  $H_1$ , which is the negation of (5), and under a sequence of local alternatives that converges to  $H_0$  at the parametric rate  $n^{-1/2}$ ,  $n$  being the sample size. In addition, we show that critical values can be computed with the assistance of a multiplier-type bootstrap that is easy to implement.

#### 3.1 Asymptotic null distribution

The asymptotic distributions of our tests are the limiting distributions of continuous functionals of  $\hat{R}_{w,n}^p(\cdot)$  under  $H_0$ . To derive the asymptotic results, we adopt the following notation. For a generic set  $\mathcal{G}$ , let  $l^\infty(\mathcal{G})$  be the Banach space of all uniformly bounded real functions on  $\mathcal{G}$  equipped with the uniform metric  $\|f\|_{\mathcal{G}} \equiv \sup_{z \in \mathcal{G}} |f(z)|$ . We study the weak convergence of  $\hat{R}_{w,n}^p(\cdot)$  and related processes as elements of  $l^\infty(\Pi)$ , where  $\Pi \equiv [0, 1]$ . Let  $\Rightarrow$  denote weak convergence on  $(l^\infty(\Pi), \mathcal{B}_\infty)$  in the sense of J. Hoffmann-Jørgensen, where  $\mathcal{B}_\infty$  denotes the corresponding Borel  $\sigma$ -algebra - see e.g. Definition 1.3.3 in [van der Vaart and Wellner \(1996\)](#).

We assume the following regularity conditions.

**Assumption 1** (i) The parameter space  $\Theta$  is a compact subset of  $\mathbb{R}^k$ ; (ii) the true parameter  $\theta_0$  belongs to the interior of  $\Theta$ ; and (iii)  $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$ .

**Assumption 2** (i)  $q(x, \theta)$  is twice continuously differentiable at each  $\theta$  in a neighborhood of  $\theta_0$ . Denote

$$g(x, \theta) = \frac{\partial q(x, \theta)}{\partial \theta} = (g_1(x, \theta), \dots, g_k(x, \theta))';$$

(ii) for all  $\theta \in \Theta$ , there exists an integrable function  $M(x)$  such that  $\max_{1 \leq j \leq k} |g_j(x, \theta)| \leq$

$M(x)$ ; and (iii) the matrix  $\Delta(\theta) \equiv \mathbb{E}[g(X, \theta)g'(X, \theta)]$  is non-singular at each  $\theta$  in a neighborhood of  $\theta_0$ .

**Assumption 3** *The parametric family of weight functions  $\mathcal{W} \equiv \{w(q(\cdot, \theta), u) : \theta \times u \in \Theta \times \Pi\}$  is such that the equivalence between (3) and (5) holds. Furthermore,  $\mathcal{W}$  is a Donsker class of functions.*

Assumption 1 is a standard one. Under some standard moment conditions, it is satisfied for ML, NLLS and GMM estimators, for example. Assumption 2 is a condition concerning the degree of smoothness of the propensity score  $q(x, \theta)$ , and is satisfied for standard parametric models such as the Probit and the Logit specifications. Assumption 3 imposes some additional regularity conditions on the family of weights  $w$  and allows for the use of the most popular weight functions, including in particular the indicator and the exponential weights,  $w(q(\cdot, \theta), u) = 1\{q(\cdot, \theta) \leq u\}$ , and  $w(q(\cdot, \theta), u) = \exp(iuq(\cdot, \theta))$ , respectively.

Next, we derive the asymptotic distribution of the projection-based empirical process  $\hat{R}_{w,n}^p(\cdot)$  under  $H_0$ . We do this in two steps. First, we show that, under  $H_0$ ,  $\hat{R}_{w,n}^p(\cdot)$  is asymptotically equivalent, with respect to the supremum norm on  $\Pi$ , to the process

$$R_{w,n0}^p(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) \mathcal{P}w(q(X_i, \theta_0), u), \quad (14)$$

where  $\mathcal{P}w(q(X, \theta), u)$  is as defined in (9). From this result it follows that the weak convergence under  $H_0$  of the process  $\hat{R}_{w,n}^p(u)$  can be conveniently established from that of  $R_{w,n0}^p(u)$  given in (14). More importantly, the limiting behavior of  $\hat{R}_{w,n}^p(u)$  does not depend on  $\hat{\theta}_n$ .

**Theorem 1** *Let Assumptions 1-3 hold. Then, under  $H_0$ , we have that*

$$\sup_{u \in \Pi} \left| \hat{R}_{w,n}^p(u) - R_{w,n0}^p(u) \right| = o_p(1),$$

and

$$\hat{R}_{w,n}^p(u) \Rightarrow R_{w,\infty}^p,$$

where  $R_{w,\infty}^p$  denotes a Gaussian process with mean zero and covariance structure given

by

$$K_w^p(u_1, u_2) = \mathbb{E}[q(X, \theta_0)(1 - q(X, \theta_0))\mathcal{P}w(q(X, \theta_0), u_1)\mathcal{P}w(q(X, \theta_0), u_2)] \quad (15)$$

Theorem 1 and the continuous mapping theorem(CMT), see e.g. Theorem 1.3.6 in van der Vaart and Wellner (1996), yield the asymptotic null distributions of continuous functionals of  $\hat{R}_{w,n}^p(u)$ , including the test statistics  $CvM_n$  and  $KS_n$  given in (12) and (13), respectively.

**Corollary 1** *Under the assumptions of Theorem 1 and  $H_0$ , for any continuous functional  $\Gamma(\cdot)$ , we have*

$$\Gamma(\hat{R}_{w,n}^p) \xrightarrow{d} \Gamma(R_{w,\infty}^p).$$

Furthermore,

$$CvM_n \xrightarrow{d} CvM_\infty := \int_{\Pi} |R_{1,\infty}^p(u)|^2 dF_{\theta_0}(u),$$

where  $F_{\theta_0}(u) = \mathbb{P}(q(X, \theta_0) \leq u)$  denotes the distribution function of  $q(X, \theta_0)$ , and

$$KS_n \xrightarrow{d} KS_\infty := \sup_{u \in \Pi} |R_{1,\infty}^p(u)|.$$

Note that the integrating measure in  $CvM_n$  is a random measure, but Corollary 1 shows that the asymptotic theory is not affected by this fact. Further details can be found in the Appendix.

### 3.2 Asymptotic power against fixed and local alternatives

Now, we investigate the power properties of tests based on continuous functionals  $\Gamma(\hat{R}_{w,n}^p)$ , like  $CvM_n$  and  $KS_n$  in (12) and (13), respectively. We consider both fixed and local alternatives that converge to  $H_0$  at the parametric rate  $n^{-1/2}$ .

Next theorem analyses the asymptotic properties of our tests under the fixed alternative

$$H_1 : \mathbb{P}(\mathbb{E}[D - q(X, \theta) | q(X, \theta)] = 0) < 1, \quad \forall \theta \in \Theta. \quad (16)$$

Note that  $H_1$  is simply the negation of  $H_0$  (3).

**Theorem 2** *Suppose Assumptions 1-3 hold. Then, under the fixed alternative hypothesis  $H_1$  in (16), we have that*

$$\sup_{u \in \Pi} \left| \frac{1}{\sqrt{n}} \hat{R}_{w,n}^p(u) - \mathbb{E}[(p(X) - q(X, \theta_0)) \mathcal{P}w(q(X, \theta_0), u)] \right| = o_p(1).$$

Theorem 2 implies that the test statistic  $\Gamma(\hat{R}_{w,n}^p)$  will diverge to infinity under the alternative hypothesis  $H_1$ , because, under  $H_1$ ,  $\mathbb{E}[(p(X) - q(X, \theta_0)) \mathcal{P}w(q(X, \theta_0), \cdot)] \neq 0$  for a set with positive Lebesgue measure. Nonetheless, it is important to observe that, since our tests are based on (3), there are certain alternative hypotheses other than (16) in which our tests will have trivial power. These type of alternatives only arises when  $\mathbb{E}[D - q(X, \theta_0) | X = x] \neq 0$  for some  $x \in \mathcal{X}$ , yet  $\mathbb{E}[D - q(X, \theta_0) | q(X, \theta_0)] = 0$  a.s.. As pointed out by Shaikh et al. (2009), the only case this happens is when the conditional expectation of  $D$  given a subset of available covariates is correctly specified, but the omitted regressors have non-zero effect on the true propensity score, e.g. when  $X = (X_1, X_2)$ , and  $E[D | X_1] = q(X, \theta_0) \neq p(X) = E[D | X_1, X_2]$ . As one can see, such class of alternative hypotheses is rather exceptional. Furthermore, such cases can be circumvented by including all available covariates into the specification of  $q(X, \theta_0)$ .

Next, we study the performance of our projection-based tests under certain types of local alternatives. In particular, we derive the asymptotic distribution of  $\hat{R}_{w,n}^p$  under a sequence of alternative hypotheses converging to the null at the parametric rate  $n^{-1/2}$  given by

$$H_{1n} : E[D - q(X, \theta_0) | q(X, \theta_0)] = \frac{r(q(X, \theta_0))}{\sqrt{n}} \quad a.s. \quad (17)$$

for some  $\theta_0 \in \Theta$ . The function  $r : [0, 1] \rightarrow \mathbb{R}$  is required to satisfy the following assumption.

**Assumption 4** *The function  $r(q(X, \theta_0))$  is continuous in  $q$  a.s. and satisfies  $\mathbb{E}|r(q(X, \theta_0))| < \infty$ .*

**Theorem 3** *Suppose Assumptions 1-4 hold. Then, under the local alternatives  $H_{1n}$  given by (17), we have*

$$\hat{R}_{w,n}^p \Rightarrow R_{w,\infty}^p + \Delta_{r,w},$$

where  $R_{w,\infty}^p$  is the same Gaussian process as defined in Theorem 1, and  $\Delta_{r,w}$  is a deterministic shift function given by

$$\Delta_{r,w}(u) \equiv \mathbb{E}[r(q(X, \theta_0)) \mathcal{P}w(q(X, \theta_0), u)].$$

Note that, in general, the deterministic shift function  $\Delta_{r,w}(u) \neq 0$  for some  $u \in \Pi$ , implying that tests based on continuous even functionals of  $\hat{R}_{w,n}^p(\cdot)$  will have non-trivial power against local alternatives of the form in (17). A situation in which our tests will have trivial power against such alternatives is when directions  $r(q(X, \theta_0))$  are a linear combination of  $g(X, \theta_0)$ , that is, when  $r(q(x, \theta_0)) = \beta g(x, \theta_0)$  for some  $\beta$ .

### 3.3 Computation of critical values

From the above theorems, we see that the asymptotic distribution of continuous functionals  $\Gamma(\hat{R}_{w,n}^p)$  depend on the underlying data generating process, on the class of weight functions  $w(q, u)$  used, and of course on  $\Gamma(\cdot)$  itself. Furthermore, the complicated structure of  $K_w^p$  given in (15) does not allow for a simple representation of  $R_{w,\infty}^p$  in terms of well-known distribution-free process for which critical values are readily available. To overcome this problem, we propose to compute critical values with the assistance of a multiplier bootstrap. The proposed procedure has good theoretical and empirical properties, is straightforward to verify its asymptotic validity, computationally easy to implement, and does not require computing new parameter estimates at each bootstrap replication.

More precisely, in order to estimate the critical values, we propose to approximate the asymptotic null distribution of  $\hat{R}_{w,n}^p(u)$  by that of

$$\hat{R}_{w,n}^{p*}(u) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) \mathcal{P}_n w(q(X_i, \hat{\theta}_n), u) V_i, \quad (18)$$

where  $\{V_i\}_{i=1}^n$  is a sequence of *i.i.d.* random variables with zero mean, unit variance and bounded support, independent of the original sample  $\{(D_i, X_i')'\}_{i=1}^n$ . A popular example involves *i.i.d.* Bernoulli variates  $\{V_i\}$  with  $\mathbb{P}(V = 1 - \kappa) = \kappa/\sqrt{5}$  and  $\mathbb{P}(V = \kappa) = 1 - \kappa/\sqrt{5}$ , where  $\kappa = (\sqrt{5} + 1)/2$ , as suggested by Mammen (1993).

With  $\hat{R}_{w,n}^{p*}(u)$  at hands, the bootstrapped version of our test statistics  $\Gamma(\hat{R}_{w,n}^p)$  is

simply given by  $\Gamma\left(\hat{R}_{w,n}^{p,*}\right)$ . For instance, the bootstrapped versions of  $CvM_n$  and  $KS_n$  in (12) and (13), respectively, are given by

$$CvM_n^* = \frac{1}{n} \sum_{i=1}^n \left[ \hat{R}_{1,n}^{p,*} \left( q \left( X_i, \hat{\theta}_n \right) \right) \right]^2,$$

$$KS_n^* = \sup_{u \in \Pi} \left| \hat{R}_{1,n}^{p,*}(u) \right|.$$

The asymptotic critical values are then estimated by

$$c_{n,\alpha}^{\Gamma,*} \equiv \inf \left\{ c_\alpha \in [0, \infty) : \lim_{n \rightarrow \infty} \mathbb{P}_n^* \left\{ \Gamma \left( \hat{R}_{w,n}^{p,*} \right) > c_\alpha \right\} = \alpha \right\},$$

where  $\mathbb{P}_n^*$  means bootstrap probability, i.e. conditional on the original sample  $\{(D_i, X_i')'\}_{i=1}^n$ .

In practice,  $c_{n,\alpha}^{\Gamma,*}$  is approximated as accurately as desired by  $\left( \Gamma \left( \hat{R}_{w,n}^{p,*} \right) \right)_{B(1-\alpha)}$ , the  $B(1-\alpha)$ -th order statistic from  $B$  replicates  $\left\{ \Gamma \left( \hat{R}_{w,n}^{p,*} \right) \right\}_{l=1}^B$  of  $\Gamma \left( \hat{R}_{w,n}^{p,*} \right)$ .

The next theorem establishes the asymptotic validity of the multiplier bootstrap procedure proposed above.

**Theorem 4** *Assume Assumptions 1-3. Then*

$$\hat{R}_{w,n}^{p,*} \xRightarrow[*]{*} R_{w,\infty}^p \quad a.s.,$$

where  $R_{w,\infty}^p$  is the Gaussian process defined in Theorem 1, and  $\xRightarrow[*]{*}$  denotes the weak convergence under the bootstrap law, i.e. conditional on the original sample  $\{(D_i, X_i')'\}_{i=1}^n$ . Additionally, for any continuous functional  $\Gamma(\cdot)$ , we have  $\Gamma \left( \hat{R}_{w,n}^{p,*} \right) \xRightarrow[*]{d} \Gamma \left( R_{w,\infty}^p \right)$  a.s. under the bootstrap law.

## 4 Monte Carlo simulation study

In this section, we conduct a series of Monte Carlo experiments in order to study the finite sample properties of our proposed projection-based tests. In particular, the performance of our Cramér-von Mises and Kolmogorov-Smirnov tests  $CvM_n$  and  $KS_n$  given in (12)

and (13) is compared to the Shaikh et al. (2009) test,

$$T_n(h_n) = \sqrt{\frac{n-1}{n}} \frac{nh_n^{1/2} \hat{V}_n(h_n)}{\sqrt{\hat{\Sigma}_n(h_n)}},$$

where  $\hat{V}_n(h_n)$  is given in (4), and

$$\hat{\Sigma}_n(h_n) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \frac{1}{h_n} K^2 \left( \frac{q(X_i, \hat{\theta}_n) - q(X_j, \hat{\theta}_n)}{h_n} \right) \varepsilon_i^2(\hat{\theta}_n) \varepsilon_j^2(\hat{\theta}_n).$$

Critical values for  $CvM_n$  and  $KS_n$  are obtained using the multiplier-bootstrap procedure described in Section 3.3. For  $T_n(h_n)$ , we use the critical values from the standard normal distribution.

As in Shaikh et al. (2009), we consider samples sizes  $n$  equal to 100, 200, 400, 500, 800 and 1,000. For each design, we consider 10,000 Monte Carlo experiments. The  $\{V_i\}_{i=1}^n$  used in the bootstrap implementations are independently generated as  $V$  with  $\mathbb{P}(V = 1 - \kappa) = \kappa/\sqrt{5}$  and  $\mathbb{P}(V = \kappa) = 1 - \kappa/\sqrt{5}$ , where  $\kappa = (\sqrt{5} + 1)/2$ , as proposed by Mammen (1993). The bootstrap critical values are approximated using 1,000 replications. To compute Shaikh et al. (2009) test, we use the normal kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

The bandwidth  $h_n$  is chosen to be equal to  $cn^{-1/8}$  for  $c$  equal to 0.01, 0.05, 0.10, 0.15 and 0.50. Note that our selection of  $c$  goes beyond those considered by Shaikh et al. (2009) simulations ( $c = 0.05, 0.10$  and  $0.15$ ). We do this to assess how sensitive Shaikh et al. (2009) test may be with respect to the bandwidth  $h_n$ .



## 4.1 Simulation 1

We first consider the following data-generating processes (DGP):

$$DGP1. D^* = 1 + X_1 + X_2 - \varepsilon, \quad \varepsilon \sim N(0, 1);$$

$$DGP2. D^* = 1 + X_1 + X_2 + X_1X_2 - \varepsilon, \quad \varepsilon \sim N(0, 1);$$

$$DGP3. D^* = (1 + X_1 + X_2)^2 - \varepsilon, \quad \varepsilon \sim N(0, 1);$$

$$DGP4. D^* = 1 + X_1 + X_2 - \varepsilon, \quad \varepsilon \sim \chi_1^2;$$

$$DGP5. D^* = 1 + X_1 + X_2 - \varepsilon, \quad \varepsilon \sim U(-1, 1).$$

For each of these DGP,  $D = 1 \{D^* > 0\}$ ,  $\varepsilon \perp (X_1, X_2)$ , where  $X_1 = Z_1$ ,  $X_2 = (Z_1 + Z_2)/\sqrt{2}$ , and  $Z_1$  and  $Z_2$  are independent standard normal random variables. All the DGPs considered are taken from [Shaikh et al. \(2009\)](#).

For DGP1-DGP5, the  $H_0$  considered is

$$H_0 : \exists \theta_0 = (\beta_0, \beta_1, \beta_2)' \in \Theta : E[D|X_1, X_2] = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2) \text{ a.s.},$$

where  $\Phi(\cdot)$  is the cumulative distribution function (CDF) of the standard normal distribution. We estimate  $\theta_0$  using the Probit ML, i.e.

$$\hat{\theta}_n = \arg \max_{\alpha_0, \alpha_1, \alpha_2} \sum_{i=1}^n D_i \ln(\Phi(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i})) + (1 - D_i) \ln(1 - \Phi(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i})).$$

DGP1 falls under  $H_0$ , whereas DGP2-DGP5 fall under  $H_1$ . The simulation results are presented in Table 1. We report empirical rejections at the 5% significance level. Results for 10% and 1% significance levels are similar and available upon request.

We first analyze the size of our test. From the result of DGP1, we find that the actual finite sample size of both  $KS_n$  and  $CvM_n$  tests is very close to their nominal size, even when the sample size is as small as 100. On the other hand, we find that [Shaikh et al. \(2009\)](#)'s test are, in general, conservative, and sensitive to the choice of bandwidth. For instance, when  $c = 0.5$ , the empirical size is very close to zero even with  $n = 1,000$ . On the other hand, with  $c = 0.01$ , the empirical size of [Shaikh et al. \(2009\)](#) test is closer to the nominal value.

**Table 1:** Proportion of rejections at 5% level.

DGP	$n$	$CvM_n$	$KS_n$	$T_n(0.01)$	$T_n(0.05)$	$T_n(0.10)$	$T_n(0.15)$	$T_n(0.50)$
1	100	5.98	5.25	5.67	4.72	2.18	0.78	0.01
1	200	5.83	5.48	5.96	3.61	1.49	0.77	0.00
1	400	5.53	5.56	5.31	3.53	1.52	0.87	0.00
1	500	5.17	5.54	4.85	3.49	1.77	0.89	0.02
1	800	5.27	5.63	4.33	3.10	1.71	0.99	0.02
1	1000	5.33	5.06	4.65	3.45	1.90	1.23	0.03
2	100	60.55	40.44	7.92	16.03	18.99	16.47	0.00
2	200	93.48	80.71	18.62	49.75	59.22	59.18	1.55
2	400	99.94	99.77	57.42	92.76	96.54	97.09	52.11
2	500	100	99.98	74.34	98.34	99.43	99.50	84.27
2	800	100	100	97.05	99.99	100	100	99.92
2	1000	100	100	99.53	100	100	100	100
3	100	81.40	69.96	28.43	45.70	29.83	13.63	0.00
3	200	98.12	95.72	74.39	87.96	82.11	60.47	0.00
3	400	99.99	99.94	98.00	99.59	99.45	98.28	0.13
3	500	100	99.98	99.47	99.92	99.90	99.72	0.57
3	800	100	100	100	100	100	100	11.15
3	1000	100	100	100	100	100	100	35.51
4	100	51.61	43.81	6.72	13.38	16.43	17.03	1.20
4	200	89.75	83.02	14.11	38.40	49.07	51.78	14.61
4	400	99.82	99.49	41.41	84.66	92.19	93.43	73.08
4	500	99.99	99.95	57.42	94.87	98.03	98.46	89.72
4	800	100	100	89.97	99.93	99.99	100	99.74
4	1000	100	100	97.51	99.99	99.99	99.99	99.99
5	100	6.69	5.82	3.92	5.90	4.09	2.12	0.01
5	200	7.20	8.08	6.05	4.40	2.72	1.37	0.00
5	400	8.37	10.41	5.33	4.53	3.60	2.88	0.03
5	500	8.78	11.35	4.83	5.14	4.74	3.61	0.01
5	800	10.82	16.15	5.26	7.75	8.23	7.34	0.29
5	1000	13.23	20.01	6.32	10.95	12.16	11.29	0.83

Note: Simulations based on 10,000 Monte Carlo experiments. “ $CvM_n$ ” and “ $KS_n$ ” stands for our proposed Cramér-von Mises and Kolmogorov-Smirnov tests. “ $T_n(c)$ ” stands for [Shaikh et al. \(2009\)](#) test, with bandwidth  $h_n = cn^{-1/8}$ .

We now go on to consider the relative performance of the tests in terms of power. Our proposed  $KS_n$  and  $CvM_n$  tests performs admirably well for the DGPs given by  $DGP2 - DGP4$ , even with  $n = 100$ . In these scenarios,  $CvM_n$  performs better than  $KS_n$ . The alternative hypothesis in  $DGP5$  is harder to detect, since the only source of misspecification in  $DGP5$  is the distribution of  $\varepsilon$ , which in turn shares many features with the normal distribution, such as the symmetry around zero. In such design  $KS_n$  performs better than  $CvM_n$ . Looking at the results for [Shaikh et al. \(2009\)](#) test, we

note that bandwidth choice can play a very important role: with  $c = 0.10$ , their test has substantially more power than when  $c$  equal to 0.01, or 0.50. In fact, when  $c = 0.50$ , their test has little to no power to detect any type of alternatives when sample size is small ( $n \leq 200$ ).

Perhaps, what is more important to emphasize in terms of power is that in all alternative hypotheses and sample sizes analyzed, our projection based tests have higher power than [Shaikh et al. \(2009\)](#) test, regardless of the bandwidth choice. Such feature highlights the advantages of our tests when compared to alternative procedures.

## 4.2 Simulation 2

In this simulation, we push forward the dimensionality of the covariates to see how our and [Shaikh et al. \(2009\)](#) tests perform in scenarios with 10 continuous covariates. To investigate further this issue, we consider the following DGPs:

$$\begin{aligned}
DGP6. D^* &= 1 + \sum_{j=1}^{10} X_j - \varepsilon, \quad \varepsilon \sim N(0, 10); \\
DGP7. D^* &= 1 + \sum_{j=1}^{10} X_j - X_1 X_2 - \varepsilon, \quad \varepsilon \sim N(0, 10); \\
DGP8. D^* &= 1 + \sum_{j=1}^{10} X_j - X_1 \sum_{k=2}^5 X_k - \varepsilon, \quad \varepsilon \sim N(0, 10); \\
DGP9. D^* &= 1 + \sum_{j=1}^{10} X_j - \sum_{k=1}^5 X_k^2 - \varepsilon, \quad \varepsilon \sim N(0, 10); \\
DGP10. D^* &= 1 + \sum_{j=1}^{10} X_j - X_1 \sum_{k=2}^5 X_k - \sum_{k=1}^5 X_k^2 - \varepsilon, \quad \varepsilon \sim N(0, 10),
\end{aligned}$$

where  $X_1$  and  $X_2$  are defined as before,  $\{X_i\}_{i=3}^{10}$  are independent standard normal random variables,  $D = 1\{D^* > 0\}$ , and  $\varepsilon \perp\!\!\!\perp \underline{X}$ , with  $\underline{X} = (1, X_1, X_2, \dots, X_{10})'$ . We increase the variance of  $\varepsilon$  to avoid  $D$  having a (close to) degenerated distribution.

For  $DGP6$ - $DGP10$ , the  $H_0$  considered is

$$H_0 : \exists \theta_0 \in \Theta : E[D|\underline{X}] = \Phi(\underline{X}'\theta_0) \text{ a.s..} \quad (19)$$

We estimate  $\theta_0$  by ML. Note that  $DGP6$  falls under  $H_0$ , whereas  $DGP7$ - $DGP10$  fall

under  $H_1$ . The simulation results for  $DGP6$ - $DGP10$  are presented in Table 2.

As before, we first discuss the size properties of the tests. From the results of  $DGP6$ , we find that  $KS_n$  and  $CvM_n$  tests are oversized when  $n$  is relatively small, but as  $n$  increases, the empirical size gets very close to its nominal value. Shaikh et al. (2009) test, on the other hand, tends to be very conservative (with the exception when  $c = 0.01$ ), and sensitive to the choice bandwidth.

Next, let us discuss the relative performance of the tests in terms of power. What is clear from Table 2 is that, regardless of the sample size and bandwidth considered, Shaikh et al. (2009) test seems to have no power to detect the alternatives described in  $DGP7$ ,  $DGP9$ , and  $DGP10$ . For  $DGP8$ , the maximum power for their test is approximately 30% when  $n = 1,000$  and  $c = 0.15$ . However, when one sets  $c = 0.01$ , the power of Shaikh et al. (2009) test reduces to approximately 10%, highlighting again how important (and non-trivial) is to “appropriately” choose the bandwidth.

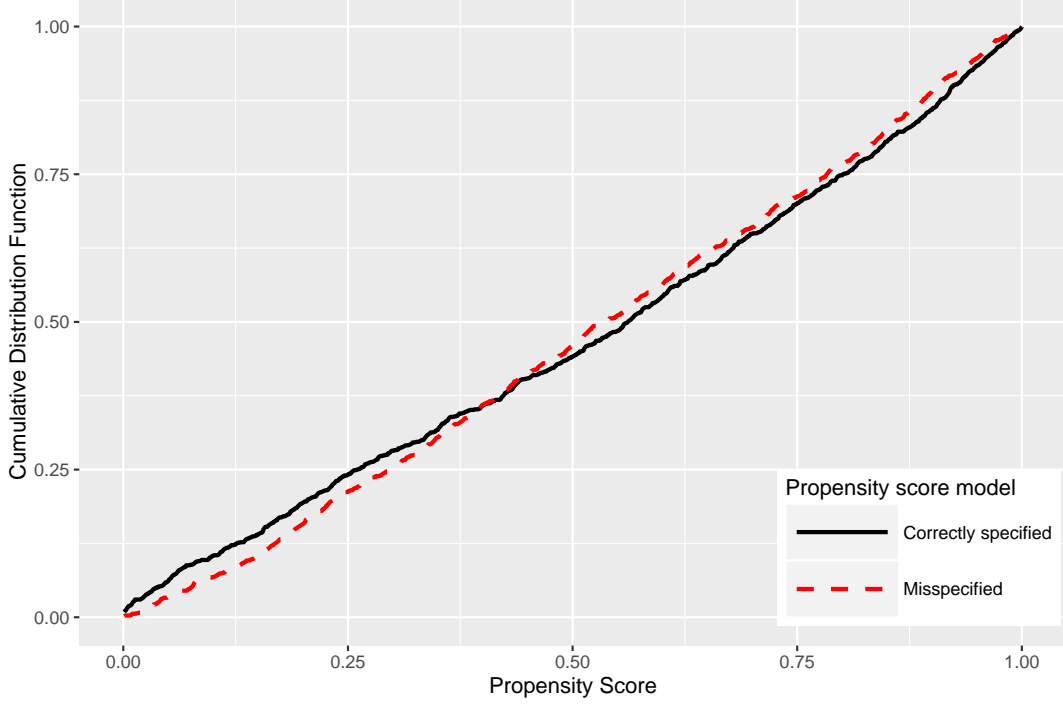
In sharp contrast with Shaikh et al. (2009) test, note that for moderately sized samples, our proposed  $KS_n$  and  $CvM_n$  tests have non-trivial power to detect all the alternatives. Perhaps, the only non-satisfactory result is related to  $DGP7$ , in which our tests achieves the highest power of appropriately 15%. Nonetheless, one must bare in mind that such alternative is hard to detect. To show this, we generate an *i.i.d.* sequence of random variables  $\{(D_i, \mathbf{X}'_i, \varepsilon_i)\}_{i=1}^{1000}$  according to the model given by  $DGP7$ , and estimate a correctly specified propensity score model which include the interaction term between  $X_1$  and  $X_2$ , and also a misspecified model that simply omit such interaction. Figure 1 displays the resulting estimates of the *CDF* of such estimated propensity score models. As one can clearly see, the two graphs lie very close to one another, suggesting that indeed alternative hypotheses like  $DGP7$  are hard to detect.

**Table 2:** Proportion of rejections at 5% level.

DGP	n	$CvM_n$	$KS_n$	$T_n(0.01)$	$T_n(0.05)$	$T_n(0.10)$	$T_n(0.15)$	$T_n(0.50)$
6	100	7.93	9.27	5.17	3.65	1.47	0.82	0.00
6	200	6.20	6.86	4.69	2.93	1.30	0.86	0.00
6	400	5.76	5.90	4.90	2.91	1.34	0.87	0.00
6	500	5.76	5.94	4.77	3.59	1.72	1.02	0.00
6	800	5.54	5.55	5.28	3.13	1.76	1.26	0.01
6	1000	5.07	5.37	4.88	3.14	1.53	0.96	0.01
7	100	8.35	8.85	5.38	3.31	1.42	0.72	0.00
7	200	7.74	7.27	5.04	3.08	1.42	0.98	0.00
7	400	8.80	8.15	4.75	3.27	2.11	1.47	0.03
7	500	9.22	8.38	4.75	3.12	2.14	1.61	0.02
7	800	11.24	9.71	4.54	3.47	2.29	1.99	0.08
7	1000	11.87	10.43	4.54	3.40	2.68	2.32	0.17
8	100	10.12	10.15	5.03	3.36	1.71	1.26	0.00
8	200	12.30	11.35	4.50	3.55	3.07	2.76	0.03
8	400	19.78	17.48	5.50	6.24	6.80	6.78	0.30
8	500	24.38	20.81	5.82	8.12	9.99	10.14	0.89
8	800	37.97	31.90	7.26	15.18	19.59	21.40	5.11
8	1000	45.01	38.51	9.03	19.99	26.83	29.71	9.94
9	100	9.94	10.04	5.79	2.49	0.95	0.46	0.00
9	200	13.33	12.25	4.74	2.45	1.32	0.84	0.00
9	400	22.65	19.46	4.58	3.10	2.29	1.64	0.02
9	500	27.56	23.72	4.68	3.39	3.04	2.22	0.01
9	800	42.53	35.25	4.93	4.88	5.04	4.56	0.03
9	1000	51.87	43.63	5.01	6.17	6.45	6.01	0.11
10	100	9.86	10.13	5.44	2.46	0.81	0.30	0.00
10	200	11.96	11.10	4.53	2.21	0.97	0.49	0.00
10	400	20.13	17.28	4.27	2.50	1.63	1.06	0.00
10	500	24.36	20.82	4.28	2.60	2.07	1.53	0.00
10	800	36.55	30.72	4.43	4.28	3.82	3.23	0.00
10	1000	44.97	38.13	4.91	4.89	4.82	4.03	0.04

Note: Simulations based on 10,000 Monte Carlo experiments. “ $CvM_n$ ” and “ $KS_n$ ” stands for our proposed Cramér-von Mises and Kolmogorov-Smirnov tests. “ $T_n(c)$ ” stands for [Shaikh et al. \(2009\)](#) test, with bandwidth  $h_n = cn^{-1/8}$ .

Overall, the simulations results analyzed show that our proposed projection-based tests performs favorably compared to [Shaikh et al. \(2009\)](#) test in terms of both power and size properties. Furthermore, our tests are fully data-driven, and do not rely on the choice of tuning parameters. Thus, we believe that our tests can be of great use in practice.



**Figure 1:** Estimates of cumulative distribution function of the propensity score from the data generating process  $D = 1\{D^* > 0\}$ , where  $D^* : 1 + \sum_{j=1}^{10} X_j - X_1 X_2 - \varepsilon$ ,  $\varepsilon \sim N(0, 10)$ . Estimated misspecified model:  $q(X, \theta) = \Phi\left(\beta_0 + \sum_{j=1}^{10} \beta_j X_j\right)$ ; estimated correctly specified model:  $q(X, \theta) = \Phi\left(\beta_0 + \sum_{j=1}^{10} \beta_j X_j + \beta_{11} X_1 X_2\right)$ .

## 5 Empirical Illustration

In this section, we provide an empirical illustration of our testing procedure. Specifically, we model the probability of having Internet access before and after the emergence of Napster (June 1999) and test whether these models are correctly specified. [Hong \(2013\)](#) uses these propensity scores as inputs into a difference-in-differences matching analysis of the effects of Napster on recorded music sales. [Hong \(2013\)](#) shows that in order to overcome compositional changes challenge, in which the treatment group (Internet users) may expand over time by including more diverse individuals, one should perform the matching based on at least two propensity scores, one for the probability of being a internet user before and one for after the emergence Napster.

The data we use is from the 1996-2002 interview surveys of the Consumer Expenditure Survey (CEX) by the US Bureau of Labor Statistics. The CEX consists of random

samples of households designed to be representative of the total US population. It contains a rich set of demographic covariates including variables related to age, education, income, appliance ownership, occupation, family composition, and region, as well as information regarding Internet access. Following [Hong \(2013\)](#), we define the Internet user group as households that either spend positive amount on computer information service (which mainly consist of Internet service fees), or were living in a college dormitory (most college dormitories are highly likely to already had Internet access in the late 1990s). The pre-Napster period is from June 1997 to May 1999 (46,124 observations), and the post-Napster period is from June 1999 to June 2001 (61,526 observations). For detailed descriptive statistics for the samples used, see [Hong \(2013\)](#).

Following [Hong \(2013\)](#) we model the propensity score of Internet access separately for the pre-Napster and post-Napster period model. We consider a standard Probit model, with a vector of covariates  $X$  that includes all variables described in [Table 3](#). This is the specification used in [Hong \(2013\)](#).

**Table 3:** Baseline covariates for the propensity score.

Variable	Description	Variable	Description
constant	intercept term	hw	1 if the household is family with husband and wife only
age	age	hwchbs	1 if husband and wife with children before school
age2	age squared	hwchis	1 if husband and wife with children in school
white	1 if white	hwchas	1 if husband and wife with children after school
black	1 if black	spchbs	1 if single parent with children before school
male	1 if male	spchis	1 if single parent with children in school
hsgrad	1 if highest education is high school graduate	retired	1 if retired
lesscol	1 if some college, less than college	headwrk	1 if the head of the household is working
colgrad	1 if college graduate	spouwrk	1 if the spouse of the household is working
tv	number of televisions	incweek1	number of weeks in a year that head worked
compute	1 if has computer	inc_hrs1	number of hours in a week that head worked
soundp	1 if has sound system	incweek2	number of weeks in a year that spouse worked
vcr	1 if has VCR	inc_hrs2	number of hours in a week that spouse worked
vehq	number of vehicles	fincreal	real final income before tax
manager	1 if occupation is equal to 1	finc2	real final income before tax squared
teacher	1 if occupation is equal to 2	owner	1 if the household owns house
prof	1 if occupation is equal to 3	renter	1 if rents house
admin	1 if occupation is equal to 4	ne	1 if resides in northwest census region
sales	1 if occupation is equal to 5 or 6	mw	1 if resides in midwest census region
tech	1 if occupation is equal to 7	west	1 if resides in west census region
service	1 if occupation is equal to 8, 9 or 10	urban	1 if resides in urban area
fam_size	family size	msa	1 if resides in metropolitan statistical area
nchle11	number of children younger than 11	ps4mil	1 if resides in area with population size over 4 million
nch1217	number of children with ages between 12 and 17	ps1mil	1 if population size between 1.2 million and 4 million
persot64	number of individuals older than 64	ps330k	1 if population size between 330 thousand and 1.2 million
single	1 if single	ps125k	1 if population size between 125 thousand and 330 thousand

For this specification, we test the null

$$H_0 : \exists \theta_0 \in \Theta : E [D - \Phi(X'\theta_0) | \Phi(X'\theta_0)] = 0 \text{ a.s.},$$

against  $H_1$ , which is simple the negation of  $H_0$ . Table 4 shows the test results for each sub-sample.  $P$ -values are based on 1,000 bootstrap draws.

At the 1% level, we see that Specification 1 is rejected for both the pre-Napster and post-Napster period. This finding suggests that the propensity scores used by Hong (2013) are misspecified, potentially raising some concerns about Hong (2013) findings. To overcome such potential concerns one could use more flexible parametric propensity score models. Alternatively, one could adopt a nonparametric approach to estimate the propensity scores, see e.g. Li et al. (2009). Given that the available sample size is relatively large and that only few covariates are continuous, the “curse of dimensionality” associated with nonparametric methods may not be severe. Thus, such an alternative seems indeed feasible.

**Table 4:** Results from specification tests.

Sub-sample		Test statistic	Bootstrap $P$ -value
Pre-Napster period	$KS_n$	1.6255	0.000
	$CvM_n$	0.7631	0.000
Post-Napster period	$KS_n$	2.9395	0.000
	$CvM_n$	2.3859	0.000

## 6 Conclusion

In this article, we have proposed new nonparametric projection-based tests for the correct specification of the propensity scores. We have shown that, in contrast to other proposals, our tests are asymptotically not sensitive to the estimation method used to estimate the propensity score under the null, and do not rely on the potentially ad hoc choice of bandwidths. We have derived the asymptotic properties of the proposed tests, and have proved that they are able to detect local alternatives converging to the null at the parametric rate, and that critical values can be easily computed via a simple multiplier bootstrap procedure. Our Monte Carlo simulation study illustrates that, for a large class of alternatives, our projection-based tests perform better in finite samples than existing tests, though there are some rather exceptional classes of alternatives in which our tests have trivial power. All these finite sample findings are in line with our asymptotic results.



Finally, our empirical application concerning the effect of Napster on recorded music sales showed the feasibility and appeal of our tests in relevant scenarios. Given that the validity of many policy evaluation procedures rely on the correct specification of the propensity score, we argue that the tests proposed in this article are important additions to the applied researcher’s toolkit.

We would like to emphasize that our proposed specification tests can also be useful in contexts other than treatment effects, in which “selectivity” plays an important role. A leading example is in estimation with missing data when missingness is random conditional on a set of covariates, see e.g. [Wooldridge \(2007\)](#). A common approach to overcome the missing data problem is to model the missingness probability with a parametric specification, and then rely on inverse probability weighted estimators. Our specification tests could then be straightforwardly applied to assess the reliability of the resulting estimator.

## Appendix: Mathematical Proofs

We provide the proofs of the main theoretical results in this appendix. Several elementary lemmas are first established. Define an auxiliary process

$$\tilde{R}_{w,n}(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) w(q(X_i, \theta_0), u).$$

The first lemma states that the process  $\hat{R}_{w,n}(u)$  in (6) is asymptotically equivalent under  $H_0$  to the process  $\tilde{R}_{w,n}(u)$  defined above.

**Lemma 1:** *Under Assumptions 1-3 and under the null hypothesis  $H_0$ , we have*

$$\sup_{u \in \Pi} \left| \hat{R}_{w,n}(u) - \tilde{R}_{w,n}(u) \right| = o_p(1).$$

**Proof of Lemma 1:** We write uniformly in  $u$

$$\begin{aligned}
& \hat{R}_{w,n}(u) - \tilde{R}_{w,n}(u) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) (w(q(X_i, \hat{\theta}_n), u) - w(q(X_i, \theta_0), u)) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i(\theta_0) - (q(X_i, \hat{\theta}_n) - q(X_i, \theta_0))) (w(q(X_i, \hat{\theta}_n), u) - w(q(X_i, \theta_0), u)) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) (w(q(X_i, \hat{\theta}_n), u) - w(q(X_i, \theta_0), u)) \\
&\quad - \sqrt{n}(\hat{\theta}_n - \theta_0)' \frac{1}{n} \sum_{i=1}^n g(X_i, \theta_0) (w(q(X_i, \hat{\theta}_n), u) - w(q(X_i, \theta_0), u)) + o_p(1) \\
&:= A_{1n} + A_{2n} + o_p(1),
\end{aligned}$$

where the second to last equality follows by the first order Taylor expansion and Assumptions 1 and 2.

Define the process

$$\alpha_n(u, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta) w(q(X_i, \theta), u).$$

Since under  $H_0$  the  $\varepsilon$ 's are centered conditionally on  $X$ 's,  $\alpha_n(u, \theta)$  has *i.i.d.* centered summands. Clearly, the first term  $A_{1n}$  can be expressed as  $\alpha_n(u, \hat{\theta}_n) - \alpha_n(u, \theta_0)$ . Under Assumption 3,  $\alpha_n(\cdot, \cdot)$  is asymptotically equicontinuous, see e.g. [van der Vaart and Wellner \(1996\)](#). Since  $\hat{\theta}_n \rightarrow_p \theta_0$  by Assumptions 1(iii),  $A_{1n} \rightarrow 0$  in probability uniformly in  $u$ .

For the second term  $A_{2n}$ , since by Assumption 1(iii),  $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$ , it remains to show that, uniformly in  $u$ ,

$$\frac{1}{n} \sum_{i=1}^n g(X_i, \theta_0) (w(q(X_i, \hat{\theta}_n), u) - w(q(X_i, \theta_0), u)) \rightarrow_p 0.$$

However, this follows straightforwardly from the uniform convergence of

$$\frac{1}{n} \sum_{i=1}^n g(X_i, \theta_0) w(q(X_i, \theta), u)$$

in  $u$  and  $\theta$  together with the continuity of its limits.  $\square$

With the help of Lemma 1, the next lemma establishes the asymptotic representation

of the process  $\hat{R}_{w,n}(u)$ .

**Lemma 2:** *Under Assumptions 1-3 and under the null hypothesis  $H_0$ , we have*

$$\sup_{u \in \Pi} \left| \hat{R}_{w,n}(u) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) w(q(X_i, \theta_0), u) + \sqrt{n}(\hat{\theta}_n - \theta_0)' G_w(u, \theta_0) \right| = o_p(1),$$

where  $G_w(u, \theta) = E[g(X, \theta) w(q(X, \theta), u)]$ .

**Proof of Lemma 2:** From Lemma 1, we have uniformly in  $u$

$$\begin{aligned} \hat{R}_{w,n}(u) &= \tilde{R}_{w,n}(u) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) w(q(X_i, \theta_0), u) - \frac{1}{\sqrt{n}} \sum_{i=1}^n (q(X_i, \hat{\theta}_n) - q(X_i, \theta_0)) w(q(X_i, \theta_0), u) + o_p(1). \end{aligned}$$

By the Mean Value Theorem (MVT) and Assumption 1, the second term in the previous expression is

$$\begin{aligned} & - \sqrt{n}(\hat{\theta}_n - \theta_0)' \frac{1}{n} \sum_{i=1}^n \frac{\partial q(X_i, \tilde{\theta}_n)}{\partial \theta} w(q(X_i, \theta_0), u) \\ &= - \sqrt{n}(\hat{\theta}_n - \theta_0)' E[g(X, \theta_0) w(q(X, \theta_0), u)] + o_p(1), \end{aligned}$$

with  $\tilde{\theta}_n$  between  $\hat{\theta}_n$  and  $\theta_0$ , where the latter equality follows by the uniform law of large numbers (ULLN) of [Newey and McFadden \(1994\)](#), Lemma 2.4. This finishes the proof of Lemma 2.  $\square$

Define the following quantity

$$\hat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) g(X_i, \hat{\theta}_n).$$

**Lemma 3:** *Under Assumptions 1-3 and under the null hypothesis  $H_0$ , we have*

$$\hat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) g(X_i, \theta_0) - \Delta(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1),$$

where  $\Delta(\theta) = E[g(X, \theta) g'(X, \theta)]$ .

**Proof of Lemma 3:** We can rewrite

$$\begin{aligned}
\hat{S}_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) g(X_i, \theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i(\hat{\theta}_n) - \varepsilon_i(\theta_0)) g(X_i, \theta_0) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) (g(X_i, \hat{\theta}_n) - g(X_i, \theta_0)) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i(\hat{\theta}_n) - \varepsilon_i(\theta_0)) (g(X_i, \hat{\theta}_n) - g(X_i, \theta_0)) \\
&:= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) g(X_i, \theta_0) + C_{1n} + C_{2n} + C_{3n}.
\end{aligned}$$

We first show  $C_{1n} = -\sqrt{n}(\hat{\theta}_n - \theta_0)' \Delta(\theta_0) + o_p(1)$ . Note that

$$\begin{aligned}
C_{1n} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^n (q(X_i, \hat{\theta}_n) - q(X_i, \theta_0)) g(X_i, \theta_0) \\
&= -\frac{1}{n} \sum_{i=1}^n g(X_i, \theta_0) \frac{\partial q(X_i, \tilde{\theta}_n)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= -E[g(X, \theta_0) g'(X, \theta_0)] \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1),
\end{aligned}$$

with  $\tilde{\theta}_n$  between  $\hat{\theta}_n$  and  $\theta_0$ , where the second to last equality follows by the MVT, and the last equality follows from the ULLN of [Newey and McFadden \(1994\)](#), Lemma 2.4.

It remains to show that both  $C_{2n}$  and  $C_{3n}$  are asymptotically negligible. Note that

$$\begin{aligned}
C_{2n} &= \sqrt{n}(\hat{\theta}_n - \theta_0)' \frac{1}{n} \sum_{i=1}^n \varepsilon_i(\theta_0) \frac{\partial g(X_i, \tilde{\theta}_n)}{\partial \theta} \\
&= \sqrt{n}(\hat{\theta}_n - \theta_0)' E \left[ \varepsilon(\theta_0) \frac{\partial g(X, \theta_0)}{\partial \theta} \right] + o_p(1) \\
&= o_p(1),
\end{aligned}$$

where the first equality follows by MVT, the second equality by ULLN of [Newey and McFadden \(1994\)](#), and the last equation by Assumptions 1 and 2 as well as the law of iterated expectations under  $H_0$ .

On the other hand, for the term  $C_{3n}$ , we get

$$\begin{aligned}
\sqrt{n}C_{3n} &= -\sqrt{n}(\hat{\theta}_n - \theta_0)' \frac{1}{n} \sum_{i=1}^n \frac{\partial q(X_i, \tilde{\theta}_n)}{\partial \theta} \frac{\partial g(X_i, \tilde{\theta}_n)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= -\sqrt{n}(\hat{\theta}_n - \theta_0)' E \left[ g(X, \theta_0) \frac{\partial g(X, \theta_0)}{\partial \theta'} \right] \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\
&= O_p(1),
\end{aligned}$$

following similar arguments in proving the negligibility of  $C_{2n}$ . Hence  $C_{3n} = O_p(n^{-1/2}) = o_p(1)$ . This ends the proof of Lemma 3.  $\square$

The next two lemmas establish the uniform convergence of  $G_{n,w}(u, \hat{\theta}_n)$  and  $\Delta_n^{-1}(\hat{\theta}_n)$  to  $G_w(u, \theta_0)$  and  $\Delta^{-1}(\theta_0)$ , respectively.

**Lemma 4:** *Under Assumptions 1-3, we have*

$$\sup_{u \in \Pi} |G_{n,w}(u, \hat{\theta}_n) - G_w(u, \theta_0)| = o_p(1).$$

**Proof of Lemma 4:** The proof follows directly from the ULLN of [Newey and McFadden \(1994\)](#).  $\square$

**Lemma 5:** *Under Assumptions 1-2, we have*

$$\Delta_n^{-1}(\hat{\theta}_n) = \Delta^{-1}(\theta_0) + o_p(1).$$

**Proof of Lemma 5:** The proof follows from the ULLN of [Newey and McFadden \(1994\)](#) and the continuous mapping theorem.  $\square$

**Proof of Theorem 1:** By straightforward decomposition, we have

$$\begin{aligned} \hat{R}_{w,n}^p(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) \left( w(q(X_i, \hat{\theta}_n), u) - g'(X_i, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) G_{n,w}(u, \hat{\theta}_n) \right) \\ &= \hat{R}_{w,n}(u) - G'_{n,w}(u, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) g(X_i, \hat{\theta}_n) \\ &:= \hat{R}_{w,n}(u) - G'_{n,w}(u, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) \hat{S}_n. \end{aligned}$$

By Lemmas 2-5, we have that

$$\begin{aligned} \hat{R}_{w,n}^p(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) w(q(X_i, \theta_0), u) - G'_w(u, \theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0) \\ &\quad - G'_w(u, \theta_0) \Delta^{-1}(\theta_0) \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) g(X_i, \theta_0) - \Delta(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0) \right] + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) (w(q(X_i, \theta_0), u) - G'_w(u, \theta_0) \Delta^{-1}(\theta_0) g(X_i, \theta_0)) + o_p(1) \\ &= R_{w,n0}^p(u) + o_p(1) \end{aligned}$$

uniformly in  $u \in \Pi$ . This ends the proof of Theorem 1.  $\square$

**Proof of Corollary 1:** The weak convergence of the empirical process  $\hat{R}_{w,n}^p(u)$  and the continuous mapping theorem ensure the convergence of  $\Gamma(\hat{R}_{w,n}^p)$  to  $\Gamma(R_{w,\infty}^p)$  for any continuous functional  $\Gamma(\cdot)$  and in particular that of  $KS_n$  to  $KS_\infty$ .

For the test statistic  $CvM_n$ , we will prove that

$$\int_{\Pi} \left| \hat{R}_{1,n}^p(u) \right|^2 F_n(du) \xrightarrow{d} \int_{\Pi} \left| R_{1,\infty}^p(u) \right|^2 F_{\theta_0}(du).$$

The weak convergence of the processes  $\hat{R}_{1,n}^p(u)$  and  $\sqrt{n}(F_n(u) - F_{\theta_0}(u))$  and the Skorohod construction (see [Serfling, 1980](#)), yield

$$\sup_u \left| \hat{R}_{1,n}^p(u) - R_{1,\infty}^p(u) \right| \rightarrow_{a.s.} 0, \quad (20)$$

and

$$\sup_u |F_n(u) - F_{\theta_0}(u)| \rightarrow_{a.s.} 0. \quad (21)$$

Now write

$$\begin{aligned} \left| \int_{\Pi} \left| \hat{R}_{1,n}^p(u) \right|^2 F_n(du) - \int_{\Pi} \left| R_{1,\infty}^p(u) \right|^2 F_{\theta_0}(du) \right| &\leq \left| \int_{\Pi} \left( \left| \hat{R}_{1,n}^p(u) \right|^2 - \left| R_{1,\infty}^p(u) \right|^2 \right) F_n(du) \right| \\ &\quad + \left| \int_{\Pi} \left| R_{1,\infty}^p(u) \right|^2 (F_n(du) - F_{\theta_0}(du)) \right| \end{aligned}$$

The first term of the right-hand side of the above inequality is  $o(1)$  a.s. due to (20). The trajectories of the limit process  $R_{1,\infty}^p(u)$  are bounded and continuous almost surely. Then, by applying Helly-Bray Theorem (see p.97 in [Rao, 1965](#)) to each of these trajectories and taking into account (21), we obtain  $\left| \int_{\Pi} \left| R_{1,\infty}^p(u) \right|^2 (F_n(du) - F_{\theta_0}(du)) \right| \rightarrow_{a.s.} 0$ . This concludes the proof of Corollary 1.  $\square$

**Proof of Theorem 2:** Under Assumptions 1-3, uniformly in  $u \in \Pi$ ,

$$\begin{aligned} &\sup_{u \in \Pi} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \varepsilon_i(\hat{\theta}_n) \mathcal{P}_n w(q(X_i, \hat{\theta}_n), u) - E[\varepsilon(\theta_0) \mathcal{P} w(q(X, \theta_0), u)] \right\} \right| \\ &= \sup_{u \in \Pi} \left| \frac{1}{\sqrt{n}} \hat{R}_{w,n}^p(u) - \mathbb{E}[(p(X) - q(X, \theta_0)) \mathcal{P} w(q(X, \theta_0), u)] \right| \\ &= o_p(1) \end{aligned}$$

by ULLN of [Newey and McFadden \(1994\)](#) and similar arguments in proving Lemmas 1, 4 and 5.  $\square$

**Proof of Theorem 3:** Note that under the local alternatives  $H_{1n}$  in (17), we have that uniformly in  $u \in \Pi$ :

$$\begin{aligned}
\hat{R}_{w,n}^p(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \varepsilon_i(\hat{\theta}_n) - \frac{r(q(X_i, \hat{\theta}_n))}{\sqrt{n}} \right) \mathcal{P}_n w(q(X_i, \hat{\theta}_n), u) \\
&\quad + \frac{1}{n} \sum_{i=1}^n r(q(X_i, \hat{\theta}_n)) \mathcal{P}_n w(q(X_i, \hat{\theta}_n), u) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \varepsilon_i(\theta_0) - \frac{r(q(X_i, \theta_0))}{\sqrt{n}} \right) \mathcal{P} w(q(X_i, \theta_0), u) \\
&\quad + E[r(q(X, \theta_0)) \mathcal{P} w(q(X, \theta_0), u)] + o_p(1) \\
&:= R_{w,n1}^p(u) + \Delta_{r,w}(u) + o_p(1) \\
&\Rightarrow R_{w,\infty}^p + \Delta_{r,w},
\end{aligned}$$

where the second equality follows by same similar arguments in proving Theorem 1 and by ULLN. Since that  $\varepsilon_i(\theta_0) - n^{-1/2}r(q(X_i, \theta_0))$  is a zero mean martingale difference in this local alternative framework, we can apply the functional central limit theorem to  $R_{w,n1}^p(u)$ , just as we applied it to  $R_{w,n0}^p(u)$  defined in (14), leading to  $R_{w,n1}^p(u) \Rightarrow R_{w,\infty}^p$ .

The last step follows and we finish the proof of Theorem 3.  $\square$

**Proof of Theorem 4:** As in Theorem 1, we have the following decomposition:

$$\begin{aligned}
\hat{R}_{w,n}^{p*}(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) \left( w(q(X_i, \hat{\theta}_n), u) - g'(X_i, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) G_{n,w}(u, \hat{\theta}_n) \right) V_i \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) w(q(X_i, \hat{\theta}_n), u) V_i - G'_{n,w}(u, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\hat{\theta}_n) g(X_i, \hat{\theta}_n) V_i \\
&:= \hat{R}_{w,n}^*(u) - G'_{n,w}(u, \hat{\theta}_n) \Delta_n^{-1}(\hat{\theta}_n) \hat{S}_n^*.
\end{aligned}$$

By Assumption 3, it follows from a stochastic equicontinuity argument and the consistency of  $\hat{\theta}_n$  that, uniformly in  $u \in \Pi$ ,

$$\hat{R}_{w,n}^*(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) w(q(X_i, \theta_0), u) V_i + o_p(1),$$

and

$$\hat{S}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) g(X_i, \theta_0) V_i + o_p(1).$$

Thus, uniformly in  $u$ ,

$$\begin{aligned}
\hat{R}_{w,n}^{p*}(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) (w(q(X_i, \theta_0), u) - G'_w(u, \theta_0) \Delta^{-1}(\theta_0) g(X_i, \theta_0)) V_i + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(\theta_0) \mathcal{P}w(q(X_i, \theta_0), u) V_i + o_p(1) \\
&:= R_{w,n0}^{p*}(u) + o_p(1),
\end{aligned}$$

leading to the bootstrapped version of  $R_{w,n0}^p(u)$  in (14). The rest of the proof follows from the multiplier central limit theorem applied to  $R_{w,n0}^{p*}(u)$ ; see van der Vaart and Wellner (1996, Theorem 2.9.2, p.179), and the continuous mapping theorem.  $\square$



## References

- Abadie, A., and Imbens, G. W. (2006), “Large sample properties of matching estimators for average treatment effects,” *Econometrica*, 74(1), 235–267.
- Abadie, A., and Imbens, G. W. (2016), “Matching on the estimated propensity score,” *Econometrica*, 84(2), 781–807.
- Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006), “Tailor-made tests for goodness of fit to semiparametric hypotheses,” *Annals of Statistics*, 34(2), 721–741.
- Bierens, H. J. (1982), “Consistent model specification tests,” *Journal of Econometrics*, 20(1982), 105–134.
- Bierens, H. J. (1990), “A consistent conditional moment test of functional form,” *Econometrica*, 58(6), 1443–1458.
- Bierens, H. J., and Ploberger, W. (1997), “Asymptotic theory of integrated conditional moment tests,” *Econometrica*, 65(5), 1129–1151.
- Busso, M., Dinardo, J., and McCrary, J. (2014), “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators,” *The Review of Economics and Statistics*, 96(5), 885–895.
- Dehejia, R., and Wahba, S. (2002), “Propensity score-matching methods for nonexperimental causal studies,” *The Review of Economics and Statistics*, 84(1), 151–161.
- Delgado, M. A., and Stute, W. (2008), “Distribution-free specification tests of conditional models,” *Journal of Econometrics*, 143(1), 37–55.
- Donald, S. G., and Hsu, Y.-C. (2014), “Estimation and inference for distribution functions and quantile functions in treatment effect models,” *Journal of Econometrics*, 178(3), 383–397.
- Escanciano, J. C. (2006a), “A consistent diagnostic test for regression models using projections,” *Econometric Theory*, 22, 1030–1051.
- Escanciano, J. C. (2006b), “Goodness-of-Fit Tests for Linear and Nonlinear Time Series Models,” *Journal of the American Statistical Association*, 101(474), 531–541.
- Escanciano, J. C. (2007), “Model Checks using Residual Marked Empirical Processes,” *Statistica Sinica*, 17, 115–138.
- Escanciano, J. C. (2009), “On the Lack of Power of Omnibus Specification Tests,” *Econometric Theory*, 25(01), 162.
- Escanciano, J. C., and Goh, S. C. (2014), “Specification analysis of linear quantile models,” *Journal of Econometrics*, 178, 495–507.
- Fan, Y., and Li, Q. (1996), “Consistent model specification tests: omitted variables and semiparametric functional forms,” *Econometrica*, 64(4), 865–890.
- Firpo, S. (2007), “Efficient semiparametric estimation of quantile treatment effects,” *Econometrica*, 75(1), 259–276.
- Frölich, M. (2004), “Finite-sample properties of propensity-score matching and weighting estimators,” *The Review of Economics and Statistics*, 86(1), 77–90.
- González-Manteiga, W., and Crujeiras, R. M. (2013), “An updated review of Goodness-of-Fit tests for regression models,” *Test*, 22(3), 361–411.

- Hahn, J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66(2), 315–331.
- Hardle, W., and Mammen, E. (1993), “Comparing nonparametric versus parametric regression fits,” *The Annals of Statistics*, 21(4), 1926–1947.
- Heckman, J. J., Ichimura, H., and Todd, P. (1997), “Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme,” *The Review of Economic Studies*, 64(4605-654).
- Heckman, J. J., and Vytlacil, E. J. (2007), “Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation,” *Handbook of Econometrics*, 6B(70), 4779–4874.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003), “Efficient estimation of average treatment effects using the estimated propensity score,” *Econometrica*, 71(4), 1161–1189.
- Hong, S.-H. (2013), “Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes,” *Journal of Applied Econometrics*, 28(2), 297–324.
- Huber, M., Lechner, M., and Wunsch, C. (2013), “The performance of estimators based on the propensity score,” *Journal of Econometrics*, 175(1), 1–21.
- Imbens, G. W., and Wooldridge, J. M. (2009), “Recent developments in the econometrics of program evaluation,” *Journal of Economic Literature*, 47(1), 5–86.
- Janssen, A. (2000), “Global power functions of goodness of fit tests,” *Annals of Statistics*, 28(1), 239–253.
- Kang, J. D. Y., and Schafer, J. L. (2007), “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data,” *Statistical Science*, 22(4), 569–573.
- Lee, W.-S. (2013), “Propensity score matching and variations on the balancing test,” *Empirical Economics*, 44(1), 47–80.
- Li, Q., Racine, J. S., and Wooldridge, J. M. (2009), “Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data,” *Journal of Business & Economic Statistics*, 27(2), 206–223.
- Li, Q., and Wang, S. (1998), “A simple consistent bootstrap test for a parametric regression function,” *Journal of Econometrics*, 87(1), 145–165.
- Mammen, E. (1993), “Bootstrap and wild bootstrap for high dimensional linear models,” *The Annals of Statistics*, 21(1), 255–285.
- Newey, W. K., and McFadden, D. (1994), “Large sample estimation and hypothesis testing,” *Handbook of Econometrics*, 4, 2111–2245.
- Neyman, J. (1959), “Optimal Asymptotic Tests of Composite Statistical Hypotheses,” in *Probability and Statistics: The Harald Cramer Volume*, ed. U. Grenander, Stockholm:, pp. 213–234.
- Rao, C. R. (1965), *Linear Statistical Inference and its Applications*, New York: Wiley.
- Rosenbaum, P. R. (1987), “Model-Based Direct Adjustment,” *Journal of the American Statistical Association*, 82(398), 387–394.

- Rosenbaum, P. R., and Rubin, D. B. (1983), “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70(1), 41–55.
- Rosenbaum, P. R., and Rubin, D. B. (1985), “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score,” *The American Statistician*, 39(1), 33–38.
- Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.
- Shaikh, A. M., Simonsen, M., Vytlačil, E. J., and Yildiz, N. (2009), “A specification test for the propensity score using its distribution conditional on participation,” *Journal of Econometrics*, 151(1), 33–46.
- Smith, J. A., and Todd, P. E. (2005), “Does matching overcome LaLonde’s critique of nonexperimental estimators?,” *Journal of Econometrics*, 125, 305–353.
- Stinchcombe, M. B., and White, H. (1998), “Consistent specification testing with nuisance parameters present only under the alternative,” *Econometric theory*, 14, 295–325.
- Stute, W. (1997), “Nonparametric model checks for regression,” *The Annals of Statistics*, 25(2), 613–641.
- Stute, W., González-Manteiga, W., and Quindimil, M. P. (1998), “Bootstrap Approximations in Model Checks for Regression,” *Journal of the American Statistical Association*, 93(441), 141–149.
- Stute, W., and Zhu, L.-X. (2002), “Model Checks for Generalized Linear Models,” *Scandinavian Journal of Statistics*, 29(3), 535–545.
- van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer.
- Wooldridge, J. M. (2007), “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics*, 141(2), 1281–1301.
- Zheng, J. X. (1996), “A consistent test of functional form via nonparametric estimation techniques,” *Journal of Econometrics*, 75, 263–289.